## Claims

[c1] 1.A distributed monitor and control engine comprising:

a service-level-objective (SLO) agent, receiving measurements of an SLO objective for a web service to a web user, the measurements of the SLO objective indicating service quality for the web user accessing the web service at a web site, the SLO agent for adjusting resources at the web site to improve the measurements of the SLO objective;

a service agent, coupled to the SLO agent, for monitoring and controlling one or more tiers at the web site, wherein a request from the web user passes through a plurality of tiers, each tier having a plurality of service components each capable of performing a tier service for the request, the request being processed by performing a series of tier services of different tiers; and

local agents, running on nodes containing the service components, each local agent for monitoring status of a service component and for adjusting local computing resources available to the service component in response to commands from the service agent, each local agent reporting status to the service agent,

wherein the SLO agent uses the service agent and local agents to adjust resources at the web site to improve measurements of the SLO objective.

[c2] 2.The distributed monitor and control engine of claim 1 wherein an availability SLO fails when all service components fail on any one of the plurality of tiers, the web service becoming unavailable when one tier of the plurality of tiers has no available service components for performing the tier service;

wherein the SLO agent instructs the service agent to replicate a service component for a failing tier to another node in response to the availability SLO failing,

whereby service components for the failing tier are replicated to improve the availability SLO.

[c3] 3.The distributed monitor and control engine of claim 2 wherein when a

performance SLO fails, the SLO agent sends a message to the service agent, the service agent instructs one or more local agents to increase local computing resources, or the service agent replicates a service component to increase performance.

[c4] 4.The distributed monitor and control engine of claim 3 further comprising:

node monitors, coupled to report to the service agent when a node containing a service component fails,

whereby nodes are monitored by the node monitors and by the local agents.

[c5] 5.The distributed monitor and control engine of claim 2 wherein the plurality of tiers comprises at least three of the following tiers: a firewall tier, a web-server tier, an application-server tier, and a database-server tier, wherein service components for the web-server tier comprise web servers, wherein service components for the application-server tier comprise web applications, and wherein service components for the database-server tier comprise database servers.

[c6] 6.The distributed monitor and control engine of claim 2 further comprising:

a configuration manager with a user interface to an administrative user for the web site, the configuration manager receiving tier-configuration, service-configuration, and SLO information from the administrative user;

configuration storage, coupled to the SLO agent, for storing the tier-configuration, service-configuration, and SLO information from the configuration manager;

wherein the SLO agent compares a goal in the SLO information to the measurements of the SLO objective received to determine when to adjust resources at the web site to improve the measurements of the SLO objective.

[c7] 7.The distributed monitor and control engine of claim 6 wherein the tier-configuration information includes a list or primary servers and a list of backup servers for running the service component for the tier service;

wherein the service-configuration information includes a list of tiers performing tier services for the service to the web user;

wherein the SLO information includes a name of a service for the SLO, a goal, and an action to execute when the goal is not met.

[c8] 8.The distributed monitor and control engine of claim 7 wherein the service agent stores a subset of the tier-configuration, service-configuration, and SLO information stored by the configuration storage for the SLO agent, the subset being for tiers controlled by the service agent.

[c9] 9.The distributed monitor and control engine of claim 8 wherein the service agent comprises a plurality of service agents distributed about the web site, each service agent for monitoring and controlling a different tier at the web site, each service agent coupled to local agents for one tier.

[c10] 10.A computer-implemented method for monitoring and controlling a web site to meet a service-level objective (SLO) of a service having multiple tiers of service components, the method comprising:

when a SLO agent determines that an availability SLO is not being met:

commanding a service agent for a failing tier to replicate a service component for the failing tier that is below a tier-performance baseline and causing the SLO to not be met to increase a number of service components for the failing tier; and

sending an alarm from the service agent to the SLO agent indicating an action taken;

when a SLO agent determines that a performance SLO is not being met:

sending a message from the SLO agent to a service agent for a low-performing tier;

sending a command from the service agent to a local agent running a service component for the low-performing tier;

the local agent attempting to shift resources to the service component for the low-performing tier from lower-priority services running on a local node controlled by the local agent;

when the local agent is not able to shift resources, replicating the service component to a target node to increase a number of service components for

the low-performing tier; and

sending an alarm signal from the service agent to the SLO agent to report an action taken,

whereby availability and performance SLO violations are acted on by the SLO agent instructing the service and local agents to shift resources or replicate service components of a tier causing the violation.

[c11] 11.The computer-implemented method of claim 10 wherein replicating the service component comprises:

searching for a target node with sufficient resources to execute the service component;

replicating the service component to the target node.

[c12] 12.The computer-implemented method of claim 11 further comprising:

when the local agent is coupled to a local resource manager and the performance SLO is not being met, using the local resource manager to shift resources to the service component from lower-priority services running on a local node controlled by the local resource manager and the local agent.

[c13] 13.The computer-implemented method of claim 12 when a local agent signals to the service agent that a service component has failed:

the service agent comparing a maximum number of allowed restarts in a configuration to a current number of restarts for the service component;

when the current number of restarts exceeds the maximum number of allowed restarts, sending a message from the service agent to the SLO agent indicating a SLO availability violation;

when the current number of restarts does not exceed the maximum number, the service agent causing the local agent to execute a stop script to stop execution of the service component and a start script to re-initiate execution of the service component.

[c14] 14.The computer-implemented method of claim 13 further comprising:

when network errors are detected when restarting a service component, restarting or reconfiguring network interfaces coupled to the service agent

before executing the start script to re-initiate the service component.

[c15] 15.The computer-implemented method of claim 14 when a node monitor signals to the service agent that a network node is no longer accessible, the service agent sending a message to the SLO agent indicating a SLO availability violation for each service component that was running on the network node that is no longer accessible.

[c16] 16.A computer-program product comprising:

a computer-usable medium having computer-readable program code means embodied therein for controlling and monitoring service-level objectives, the computer-readable program code means in the computer-program product comprising:

network connection means for transmitting and receiving external requests for a service;

first tier means for receiving and partially processing external requests for the service having a service-level objective (SLO), the first tier means having a plurality of first service components each able to partially process a request when other first service components are not operational;

second tier means for receiving and partially processing requests from the first tier means, the second tier means having a plurality of second service components each able to partially process a request when other second service components are not operational;

third tier means for receiving and partially processing requests from the second tier means, the third tier means having a plurality of third service components each able to partially process a request when other third service components are not operational;

first local agent means, running on nodes for running the first service components, for monitoring and controlling the first service components of the first tier means;

second local agent means, running on nodes for running the second service components, for monitoring and controlling the second service components of the second tier means;

third local agent means, running on nodes for running the third service components, for monitoring and controlling the third service components of the third tier means;

SLO agent means, coupled to receive SLO measurements, for comparing an SLO measurement to a goal for a service and signaling a SLO violation when the goal is not met by the SLO measurement;

first service agent means, coupled to the first local agent means, for instructing the first local agent means to adjust resources to increase performance of the first service components in response to a message from the SLO agent means signaling the SLO violation when the SLO violation is caused by the first service components of the first tier means;

second service agent means, coupled to the second local agent means, for instructing the second local agent means to adjust resources to increase performance of the second service components in response to a message from the SLO agent means signaling the SLO violation when the SLO violation is caused by the second service components of the second tier means; and

third service agent means, coupled to the third local agent means, for instructing the third local agent means to adjust resources to increase performance of the third service components in response to a message from the SLO agent means signaling the SLO violation when the SLO violation is caused by the third service components of the third tier means,

whereby multiple tiers of service components are controlled.

[c17]

17.The computer-program product of claim 16 wherein:

the first service agent means is also for replicating the first service component to other nodes in response to the SLO violation to increase availability and performance of the first service components of the first tier means;

the second service agent means is also for replicating the second service component to other nodes in response to the SLO violation to increase availability and performance of the second service components of the second tier means;

the third service agent means is also for replicating the third service component to other nodes in response to the SLO violation to increase availability and performance of the third service components of the third tier means.

[c18] 18.The computer-program product of claim 16 wherein the computer-readable program code means further comprises:
restart means, in the first, second, and third service agent means, for instructing the first, second, or third local agent means to execute a re-start script to re-start the service component.

[c19] 19.The computer-program product of claim 16 wherein the computer-readable program code means further comprises:
compare means, coupled to the restart means, for limiting a number of times restart is attempted for a node, the first, second, or third service agent means signaling the SLO agent means when restart exceeds a limited number of times.

[c20] 20.The computer-program product of claim 16 wherein the first tier means comprises a web-server tier and the first service components are web-server components;
wherein the second tier means comprises an application-server tier and the second service components are applications;
wherein the third tier means comprises a database-server tier and the third service components are database-accessing servers.